

Controlling the Vocabulary for Anatomy

Baud RH, PhD, Lovis C, MD, Rassinoux A-M, PhD, Ruch P, MS, Geissbuhler A, MD
Medical Informatics Division, University Hospitals of Geneva, Switzerland

When confronted with the representation of human anatomy, natural language processing (NLP) system designers are facing an unsolved and frequent problem: the lack of a suitable global reference. The available sources in electronic format are numerous, but none fits adequately all the constraints and needs of language analysis. These sources are usually incomplete, difficult to use or tailored to specific needs. The anatomist's or ontologist's view does not necessarily match that of the linguist. The purpose of this paper is to review most recognized sources of knowledge in anatomy usable for linguistic analysis. Their potential and limits are emphasized according to this point of view. Focus is given on the role of the consensus work of the International Federation of Associations of Anatomists (IFAA) giving the Terminologia Anatomica.

Views on anatomy

Anatomy is one of the oldest medical sciences and it is of interest for most domains of medicine. In this paper, we concentrate on macroscopic anatomy, embryology and histology, with other branches left for future discussions. Specialists of anatomy working in different domains develop multiple points of view on anatomy that can be classified in three groups, as summarized in Figure 1.

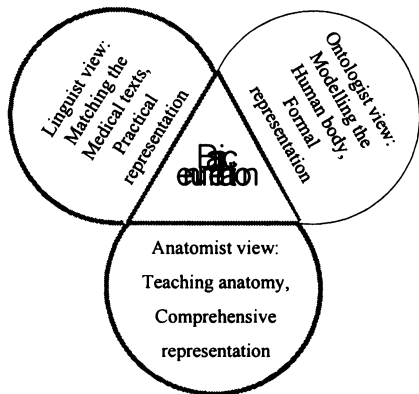


Figure 1: Different views of the anatomical reality.

The center of the triangle, intersection of the three groups contains the enumeration of the physical objects making up the human body under a universally agreed name and without ambiguity. The

different views of anatomy organize and interpret these objects according to their needs. The first view is that of the anatomist, with the major aim of teaching and understanding anatomy. The second view is that of the ontologist, seeking a logically sound representation of anatomic knowledge. The third view is that of the linguist, aiming at the recognition of anatomy terms in medical texts, irrespective of the idiosyncrasies, abbreviations and local jargon. Of course, these different views pursuing different goals are not always compatible, though large overlaps are observed. At first and before expressing any specific view, one needs a basic enumeration of all objects, in order to define the domain of representation. This is typically a task fulfilled by Terminologia Anatomica, as described hereinafter.

The birth of a standard for anatomy

In 1903, Nicolas, a French scientist founded the IFAA "whose members would use the same language for communication purpose. The vocabulary of such a language would contain a list of Latin terms to be translated into the vernacular of each nation. Each term would correspond to just one anatomical structure." [1] One century ago, everything was said, but the work was and is still to be completed.

In 1895 the Basle Nomina Anatomica (BNA) became available, but it was not universally adopted. Only in 1956 was the Nomina Anatomica (NA) edited, which later underwent different revisions. There was discussion about the addition of new sections: embryology, histology, etc. Then in 1989 the Federative Committee on Anatomical Terminology (FCAT) was created, which succeeded with the publication of Terminologia Anatomica (TA) [1] in 1998.

The TA provides in Latin and in English a terminology of some 8400 items. A hierarchical structure and a coding schema are given. The TA achieves a degree of detail, which is constant and consistent throughout all its sections. The most prominent anatomists all over the world support it. Its main value today lies in the fact that it presents a gold standard as reference terminology, not open for endless discussions, but supervised by the FCAT for structured updates. Although the TA has some

problems [2] from the point of view of knowledge representation, it is suitable for comparison with multiple sources of anatomical terms. TA is certainly more than a strict enumeration of anatomical objects. It introduces a hierarchy and carefully attempts to name each object univocally and only once. But it lacks the formal attributes and features, which would make it qualified as a representative of any specific view. In addition, we have to check that the TA is effectively available in electronic format.

The anatomist's view

The goal of the anatomist's view is the "conceptualization of the physical objects and spaces that constitute the human body at the macroscopic level of organization, specified as a machine parsable ontology that, in its human-readable form, is comprehensive to both expert and novice users of anatomical information" [3]. The main concern is teaching anatomy with increased efficacy thanks to the new computer, media tools and sources available today. A review of several problems about the best teaching practices, comparing the role of descriptive texts versus atlases is available [4].

The ontologist's view

The ontologist aims at developing a coherent model of anatomy, avoiding dependencies regarding any specific aspect. The "ontologists" can be divided into the "universalists", aiming at building the complete model of the human body and the "nominalists", whose goal is to achieve a pragmatic though exhaustive model of a local aspect of anatomy, as for example a model of the thalamus. Ontologists first try to build a hierarchical organization of anatomical concepts using differentia through *isa* links. Indeed, an ontologist is behind each anatomist and may formally be called a "taxonomist".

One of the challenges is to find a natural foundation of the hierarchy, in order to be able to present the information in a form not too different than the natural classification in body systems. A non-natural approach, though perfectly correct and eligible, would bring practical problems in its usage and therefore is not welcome. The Digital Anatomist initiative, marrying expertise of both anatomists and ontologists is certainly one of the best examples today of a pragmatic ontological approach [3].

The domain of anatomy cannot be represented using only a hierarchical structure of "Is Kind Of" (*isa*) links, due to its spatial representation in 1, 2 or 3 dimensions. In fact, a natural arborescence is to be necessarily built around the relationship "Is Part Of" (*ipo*) and its derived descendants. In addition, when

modeling the functional aspects of the anatomical concepts, a set of functional relationships is to be designed, with an "Is Branch Of" hierarchy useful for arteries, veins and nerves. There is little doubt about a final representation involving an *isa* canonical ontology and multiple interlinked hierarchies.

In order to illustrate this aspect, it is interesting to quote that both *isa* hierarchy and *ipo* hierarchy are naturally expected components of (but not limited to) a model of anatomy. Figure 2 is an example from the TA.

bones)	
bones of limb)	<i>isa</i> hierarchy
femur))
shaft of femur)	
linea aspera)	<i>ipo</i> hierarchy
lateral lip)	

Figure 2: Example of the TA hierarchy, mixing *isa* and *ipo* relationships in a regular pattern.

We see here a typical situation of a hierarchy made of *isa* links and *ipo* links. This is a necessity when modeling the anatomy. Quantitatively, more than 80% of the links of TA are *ipo* links.

Such a situation of using the *isa* and *ipo* links is not welcome by the ontologists for the following reason: the advantage of the *isa* link is to preserve the properties from a concept to its children. This is not true for the *ipo* link. For example, a property like Bones "Is Location Of" Fracture is true for Femur in Figure 2, but it is wrong for lateral lip.

In order to alleviate this situation, we have to imagine a double hierarchy. In a canonical hierarchy, all concepts are linked by *isa* links and possibly inherit attributes from their ancestors. In parallel, in an *ipo* hierarchy by body regions, the concepts are specifically interrelated when the second is a part of the first. This is the adequate solution as illustrated by the Digital Anatomist implementation [3] (see figure 3 on page 26). The same argument may be used with a branch of hierarchy for blood vessels or nerves.

There are different situations where the *ipo* relationship is used. When the partitioning clearly has a dimensional aspect, either with 0, 1, 2 or 3 dimensions, one should create descendents of *ipo* such as "Is Linear Of", "Is Surface Of" and "Is Volume Of". Also the border of a surface or of a volume should be recognized as such by specific relationships, like "Is Anterior Of", "Is Medial Of", "Is Inferior Of", etc. Other relationships are necessary for enumerating parts with functional criteria ("Is Functional Part Of"), containment aspects ("Is Contained In") or spatial neighboring ("Is Region Of"). There are numerous references on this topic [5, 6, 7],

Functional relationships are necessary for the introduction of semantic content, which is not grasped by the *isa* or *ipo* links. Typically, Cochlear nerve is a descendant of nerve, but the information that it controls the cochlea should be expressed by some functional links (and cannot be inferred by its name!).

Cochlear_nerve ISA nerve
Cochlear_nerve IsBranchOf 8_cranial_nerve
Cochlear_nerve Controls Cochlea
Cochlear_nerve HasSource Cochlear_ganglion
Cochlear_nerve HasTarget Spiral_ganglion

Figure 3: Different functional links are necessary for the description of the cochlear nerve in addition to the canonical *isa* link.

The linguist's view

Linguists are confronted with another challenge: reading medical texts of any origin and whatever the linguistic quality. Amongst the multiple annotations of a concept by several terms and their morphological variants, they have to find their way towards the perfect match, needing to solve syntactic and semantic ambiguities for this task. In addition, they have to formalize the meaning of anatomical concepts using generalization rules in order to adjust the granularity of a query to the deepness of representation of the text under scrutiny.

The linguist view best illustrates the operational aspect of working with macroscopic anatomy. Numerous surgeons, radiologists, pathologists and other physicians are permanently confronted with the precision or ambiguities of vocabulary, and need to define additional concepts for the description of situations arising from new technologies and tools.

Different anatomical sources

Whatever is the considered view on anatomy, there is an endless number of sources to be reviewed. There are lexicons [8, 9], terminologies [10], classifications [11], nomenclatures [12], thesauruses such as UMLS [13], books [14, 15], atlas [16, 17], web sites [18] and media [19, 20] with authoritative presentations of human anatomy, most of them being peer reviewed and recommended for supporting academic teaching. Our present purpose is not to provide a classification of the best sources, but to assess some characteristics and relevant elements of what is available in order to simplify the readers' choice according to their specific needs.

In order to make useful comparisons, the authors have manually compiled a map of all terms from the

different sources using a selected chapter of anatomy, under the form of a spreadsheet. Such a chapter should be reasonably complex but not too large, and the choice was the ear, not including the encephalic part of the auditory system.

The goal of this compilation, containing more than 2500 terms, is to assess the comparative coverage by the multiple sources. Most of the terms are duplicates and this hopefully means that different sources are using common terms. However, approximately 800 different terms have been counted for less than 300 objects. We are close to 3 names per object: an English term, a Latin term and one variants. Let us illustrate the complexity of the naming problem with the ganglion of the cochlear nerve. The following names are relevant: spiral ganglion, cochlear ganglion, ganglion cochleare, ganglion spirale cochleae and Corti's ganglion. And not forgetting the Eustachian tube, auditory tube, pharyngotympanic tube, tuba auditiva, and tuba auditoria.

The TA [1] has 283 terms related to the ear, not including the temporal bone and its subdivisions. As all sources, it makes a functional division between the external ear (60 terms), the middle ear (87 terms) and the internal ear (136 terms). For the middle ear, it distinctly separates the bonny labyrinth from the membranous labyrinth. In the present version, the cells are not yet classified and this would account especially for the internal ear.

The National Library of Medicine (NLM) edits the Medical Subject Headings or Mesh [10]. The purpose of this terminology is the indexation and retrieval of scientific publications in the medical domain. This is clearly a specific goal, which has influenced a number of design decisions. The number of terms for the ear in chapter A is 40. This figure clearly shows that the degree of detail of TA and other sources is not to be compared to Mesh, and that multiple anatomical terms are to be assembled under the heading of a single MeSH term.

The UMLS Meta-thesaurus [13] from the NLM is searched using TA terms for a matching concept. This process is evidently successful in most situations. The discrepancy between TA and UMLS metathesaurus comes about when the level of details of TA is deeper than UMLS and SNOMED. In our example, this is particularly visible for the inner ear, but the difference accounts for less than 10% (terms like copular caecum, reticular membrane or utricular nerve are absent of UMLS metathesaurus). In the future, the TA should be explicitly made one of the UMLS sources, including the Latin terminology.

For NLP applications, the NLM has developed the Specialist, an English lexicon of more than 130'000

entries. Such a source is expected to present a good coverage of anatomy terms. In fact we have collected a Specialist entry some 70% of all TA terms, due to the fact that numerous terms are multiword entries not intended to be present in the lexicon.

Paper lexicons are numerous and one widespread reference is certainly the Dorland [8]. Another lexicon is published in 5 different languages altogether by Elsevier [9] and amounts to 18,341 different entries, which is very useful for the discovery of alternative words. We have found that the Dorland dictionary is by far the most complete source, but it is not available as an on-line structured database for computer applications. The few missing terms of TA in this dictionary are multiword terms such as the head of malleus, where single words can be retrieved separately.

It has been thought to be of interest to retrieve all concepts present in a classification such as the International Classification of Diseases (ICD) [11]. Using a browser application on the ICD10 version, the number of different entries is 32 for the ear, giving a rate of 11 %.

The different books of anatomy and the atlas provide good matching with the TA as displayed in Table 1. The differences are due to different degrees of details in each representation. For the 3 lexicons (shaded area), the table shows the number of TA words retrieved in the lexicon.

Anatomy source	# terms	% TA	% SNOMED
Terminologia Anatomica	283	100	113.2
Medical Subject Headings	40	14.1	16.0
UMLS concepts	>260	>91.9	104.0
Specialist lexicon	>200	>70.6	80.0
Dorland's Illustrated	>270	>95.4	108.0
Elsevier's Medical Dic.	>150	>53.0	60.0
ICD version 10	32	11.3	12.8
SNOMED 3.4	250	~88.3	100
Netter Atlas		~50	~57
Sobotta Atlas		~50	~57
Gray's Anatomy		~60	~68
Moore and Dalley		~50	~57

Table 1: Number of terms found in different anatomy sources for the topic ear, expressed as percentage of TA and SNOMED

The main point coming from this table is the fact that TA, by its coverage – in addition to the authority of its authors – is de facto the reference. Alternatively, SNOMED could play this role. But, it remains true

that a few terms (~5%) are found only in TA, atlas or books, but are absent of any other electronically available sources, like vas prominens or carotid wall, (unless it is a mismatch by the authors though careful attention was given, illustrating the difficulty of the task).

Mapping words to terms

Most of anatomical terms are multi-words expressions in practically all languages. In addition, morphological variations are possible (with gender, number and case), departing from the basic representation. Latin expressions are allowed within the text as invariant strings (sometimes with rare variants).

Non-perfect scholar variants occur in different contexts, especially through omission of stop words. All these situations have to be faced by a NLP application aiming for excellent precision in the anatomy domain. Hereafter, we develop a model for the mapping of anatomy terms on free text sentences.

The author has developed a new model for word-term mapping. The UMLS solution of enumeration of all the possible variants and synonyms is not a good solution for at least 3 reasons: the enumerative task is never completed due to the endless imagination of human beings exercising their own language; the large size of the metathesaurus limits its usage; a fully developed multilingual metathesaurus is not scheduled.

Word-term mapping is achieved by the specification of one or more pivot words, one of them being necessarily present in any valid term. The pivot word is a single word in its basic form. Any pivot word may be linked to more than one term, so that a list of terms is attached through advanced pre-processing of each pivot word. The recognition of a pivot word triggers a matching algorithm working on two inputs: the immediate neighboring words of the analyzed text where the pivot word was found; the list of terms applicable to the pivot word.

The mapping process is improved using conceptual attachments. Such a technique allows the acceptance of synonyms to match word entries in a term.

Results and expectations

The first point to mention is the fact that the reconciliation of different sources is not trivial. Different terms or names are used by different authors and, at some degree of detail, it is not clear which term is referring to which concept. Even if macroscopic anatomy is basically observation of visible objects, expertise is needed to compare different sources. This problem may be confusing in

some extreme situation. Typically, the spiral membrane (TA A15.3.03.100) is what is named the basilar membrane in UMLS and SNOMED. But these 2 sources also name as the spiral membrane what is actually the vestibular membrane (TA A15.3.03.094). This later object is also called Reissner's membrane by other sources.

The second result is that the language diversity is an important factor. When reading medical texts by computer, one ideally needs some exhaustive view of the terms that are possible as expression for a given concept. The recognition of 90% of the names is not enough. Stirrup is a synonym of stapes and anvil is a other name for the incus.

The third point is the fact that adjectival forms and prefixes are quite commonly used in medical texts, in place of nouns, when naming anatomical entities. This is clearly visible in anatomy textbooks. The analysis of the words' roots, as described by the authors [21] cannot be ignored. Examples are: incudomalleal joint or utriculosaccular duct.

Conclusion

We have shown, through a simple experiment, that the variety of terms for the annotation of concepts is important, but that the full enumeration of nearly all terms is feasible at the cost of a resource-intensive search through multiple sources.

We also conclude that UMLS is the best solution today to this problem, but it is not a complete solution. More than 90% of the relevant concepts are present compared to Terminologia Anatomica, but not all the terms and their variants have been found. This problem is even bigger with other languages.

Finally, an approach with the TA provides a structured access to the concepts, hoping that this authoritative initiative should be reasonably accepted as a reference. The inclusion of the TA as a knowledge source for UMLS would without doubt be an improvement for the scientific community.

References

- [1] Terminologia Anatomica, International Anatomical Terminology, Federative Committee on Anatomical Terminology FCAT, Thieme Verlag, 1998
- [2] Rosse C. Terminologia Anatomica: Considered From the Perspective of Next-Generation Knowledge Sources. *Clinical Anatomy* 14:120-133 (2001).
- [3] Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, Brinkley JF. Motivation and Organizational Principles for Anatomical Knowledge Representation. *JAMIA* 1998;5 p 17-40.
- [4] Rosse C. Anatomy atlases. *Clinical Anatomy* 12:1999 p 293-299.
- [5] Rector AL, Gangemi A, Galeazzi E, Glowinski AJ, Rossi-Mori A. The GALEN CORE Model Schemata for Anatomy: Towards a Re-usable Application-Independent Model of Medical Concepts. *Proc of the 12th International Congress of the European Federation for Medical Informatics MIE94*, Lisbon, IOS Press, 1994 pp 229-233.
- [6] Rosse C, Shapiro LG, Brinkley JF. The Digital Anatomist Foundational Model: Principles for Defining and Structuring its Concept Domain. *AMIA Annual Fall Symposium*, 1998, pp 820-824.
- [7] Schulz EB, Price C, Brown PJB. Symbolic Anatomic knowledge representation in the Read Codes Version3: Structure and application. *J Am Med Inform Assoc*. 1997;4 pp 38-48.
- [8] Dorland's Illustrated Medical Dictionary. 28th edition. Sounders company. 1994.
- [9] Elsevier's Medical Dictionary in 5 languages. Elsevier SciencePublishing Company 2nd edition, 1975.
- [10] National Library of Medicine, Medical Subject Headings – annotated alphabetic list 1999. Bethesda MD.
- [11] International Statistical Classification of Diseases and Related Health Problems. Tenth revision. Volume 1, 2 and 3. World Health Organistion. Geneva. Switzerland. 1993 and 1997.
- [12] Coté R (ed). Systematized Nomenclature of Human and Veterinary Medicine (SNOMED International), version 3.1. American Veterinary Medical Association, 1995.
- [13] National Library of Medicine. UMLS knowledge sources, 7th ed. Bethesda MD. 1996.
- [14] Williams PL, Warwick R, Dyson M, Bannister LH. *Gray's Anatomy* 38th ed, New York: Churchill-Livinstone, 1995.
- [15] Moore KL, Dalley AF. *Clinically oriented Anatomy*. 4th edition. Lippincott, Williams & Wilkins, 1999.
- [16] Netter FH, Colacino S. *Netter Atlas of Human Anatomy*. Icon Learning Systems, LLC, 1998.
- [17] Putz R, Pabst R. *Sobotta Atlas of Human Anatomy*. Baltimore, Md, Williams & Wilkins, 1998.
- [18] Health On the Net Foundation at <http://www.hon.ch/>
- [19] Spitzer V, Ackerman MJ, Scherzinger AL, Whitlock D. The visible Human Male: A Technical Report. *JAMIA*. 1996;3: p 118-130.
- [20] Spitzer VM, Whitlock DG. The Visible Human Dataset: The Anatomical Platform for Human Simulation. *Anat Rec (New Anat)*1998;253 p 49-57.
- [21] Lovis C, Baud RH, Michel P-A, Scherrer J-R. Morphosemantic decomposition and semantic representation to allow fast and efficient natural language recognition. *J Am Med Inform Assoc* 1997; (Symposium Supplement): 873.